Un Sistema para la Identificación de Cadáveres NN en el Contexto de Búsqueda de Personas Desaparecidas

Andrea Maldonado[1], Darío Ruano[1,2], Norma Herrera[1,2], Marcelo Martínez[3]

[1] Departamento de Informática, FCFMyN, Univ.Nacional de San Luis
[2] Laboratorio de Investigación y Desarrollo en Bases de Datos , Univ. Nacional de San Luis
[3] Jefe Interino del Cuerpo Médico Forense y Criminalístico de la Tercera Circunscripción Judicial de la Provincia de Mendoza

andreamaldonadoma@gmail.com, dmruano@unsl.edu.ar, nherrera@unsl.edu.ar, drmarcelomartinez@hotmail.com

Abstract. En este trabajo abordaremos la problemática de identificación de cadáveres en el contexto de búsqueda de personas desaparecidas, usando como base el modelo de espacio métricos. El objetivo final es el desarrollo de un sistema que que permita un manejo más ágil de la información. Para ello por cada cadáver se genera un vector que contiene la información necesaria para posteriormente realizar búsquedas por similitud.
Palabras claves: Bases de Datos, Espacios Métricos, Identificación de Personas

Este trabajo se desarrolla en el marco del Trabajo Final de la Licenciatura en Ciencias de la Computación de la alumna Andrea Maldonado, dirigido por el Lic. D. Ruano y la MCs. N. Herrera. Dada la temática involucrada se cuenta con el asesoramiento del Dr. Marcelo Martínez.

1 Introducción

En la era actual, caracterizada por la evolución de las tecnologías de la información y las comunicaciones, las ciencias de la computación son transversales a la mayoría de nuestras actividades diarias, brindando las herramientas necesarias para abordar problemas complejos y contribuyendo en la búsqueda de soluciones eficientes a problemas de interés. La medicina legal y forense no escapa a esta realidad. Existen varios temas de interés en este contexto, uno de ellos es la identificación de cadáveres NN.

Dentro de los individuos que ingresan a los distintos Institutos de Medicina Forense del país, existen casos que no poseen las condiciones adecuadas para su identificación inmediata (indocumentados, en avanzado estado de descomposición, restos óseos, etc.) o sin posibilidad de identificación (fragmentos muy pequeños, restos carbonizados, etc.). Frente a esto, las instituciones deben investigar no sólo para determinar qué fue lo que sucedió (causa de la muerte) y cuándo sucedió (data de la muerte), sino también para poder dar con la identidad del cuerpo.

La importancia de identificar a las personas cuya identidad se desconoce responde no solo al derecho fundamental de todos los seres humanos de tener una identidad sino

también a numerosas razones de tipo social que van desde la necesidad de informar a los familiares de personas desaparecidas sobre la certeza de su fallecimiento hasta el hecho de evitar que personas infractoras de la ley simulen su propia muerte.

Claramente la identificación de cadáveres está directamente relacionada con la búsqueda de personas desaparecidas. En Argentina, no existe un sistema único de procesamiento para esta problemática. Cada provincia tiene su gobierno, su sistema forense y sus protocolos. Esto dificulta el proceso de identificación de cadáveres: si una persona desaparece en Chaco y aparece un cadáver similar en Chubut, no hay una forma rápida y correcta de relacionarlos. Si se contara con una bases de datos unificada en el país, cualquier investigador podría consultar cuántos hombres de entre 20 y 30 años tienen un tatuaje en el brazo derecho evitando mirar miles de expedientes que están en diferentes jurisdicciones. El Sistema Federal de Búsqueda de Personas Desaparecidas y Extraviadas (SIFEBU) es un intento de crear esta base de datos unificada pero sin tener automatizado el proceso de búsqueda.

En [6], una de las recomendaciones dadas es estandarizar el registro de cadáveres NN por categorías, evaluando la necesidad de un registro único a nivel nacional que brinde información específica sobre las características físicas e identificadoras (como huellas y muestras de ADN, por ejemplo) de cada cadáver hallado. Recomiendan además trabajar en la posibilidad de ampliar el acceso a la información que existe en los registros civiles provinciales, en los cementerios y en otros organismos nacionales como RENAPER (Registro Nacional de las Personas)

En este trabajo abordaremos esta problemática con el fin de **dar un primer paso** a un sistema federal de identificación de cadáveres en el contexto de búsqueda de personas. Debido a la complejidad de la temática, nos centraremos en la identificación de cuerpos. La investigación forense de casos que involucran la recuperación y análisis de restos óseos es un proceso complejo en el que intervienen diferentes disciplinas científicas y que no abordaremos en este trabajo. El desarrollo de una herramienta que permita un manejo más ágil de la información en el proceso de identificación de cadáveres NN, tendrá como resultado la posibilidad real y tangible de poder colaborar en una situación tan sensible como lo es identificar el cuerpo de una persona que está siendo buscada por sus seres queridos.

Lo que resta del artículo está organizado de la siguiente manera. En la Sección 2 damos el marco teórico exponiendo una reseña sobre el modelo de espacios métricos. En la Sección 3, presentamos el desarrollo realizado hasta el momento donde hemos utilizado como base el modelo de espacio métricos.. Finalizamos en la Sección 4 dando las conclusiones y el trabajo futuro.

2 El Modelo de Espacios Métricos

Las bases de datos tradicionales son construidas basándose en el concepto de búsqueda exacta: la base de datos es dividida en registros y cada registro contiene campos completamente comparables. Las consultas a la base de datos retornan todos aquellos registros cuyos campos coinciden con los aportados en tiempo de búsqueda.

Actualmente las bases de datos han incluido la capacidad de almacenar nuevos tipos de datos tales como imágenes, sonido, video, etc. Estructurar este tipo de datos en re-

gistros para adecuarlos al concepto tradicional de búsqueda exacta es difícil en muchos casos y hasta imposible si la base de datos cambia más rápido de lo que se puede estructurar (como por ejemplo la web). Aún cuando pudiera hacerse, las consultas que se pueden satisfacer con la tecnología tradicional están limitadas en variaciones de la búsqueda exacta.

Nos interesan las búsquedas en donde se puedan recuperar objetos *similares* a uno dado. Este tipo de búsqueda se conoce con el nombre de ***búsqueda por similitud***, y surge en diversas áreas; reconocimiento de voz, reconocimiento de imágenes, compresión de texto, biología computacional, son algunas de ellas.

Todas estas aplicaciones tienen algunas características comunes. Existe un universo $\mathcal{X}$ de objetos y una función de distancia $d : \mathcal{X} \times \mathcal{X} \rightarrow R^+$ que modela la similitud entre los objetos. El par $(\mathcal{X}, d)$ es llamado ***espacio métrico*** [5]. La base de datos es un conjunto $\mathcal{U} \subseteq \mathcal{X}$, el cual se preprocesa a fin de resolver búsquedas por similitud eficientemente. Se pueden mencionar tres tipos de búsquedas que normalmente se utilizan en espacios métricos [5].

Búsqueda por rango $(q, r)_d$**:** dado un elemento $q \in U$ y un radio de tolerancia r, una búsqueda por rango consiste en recuperar los objetos de la base de datos que estén a distancia a lo sumo r de q, es decir: $(q, r)_d = \{u \in U : d(q, u) \leq r\}$

Búsqueda del vecino más cercano NN(q): consiste en recuperar el (o los) elemento(s) más cercano(s) a un elemento q dado. En símbolos: $NN(q) = \{u \in U : \forall v \in U, d(q, u) \leq d(q, v)\}$

Búsqueda de los k-vecinos más cercanos NNk(q): Se busca recuperar los k elementos más cercanos a q en U. Esto significa encontrar un conjunto $A \subseteq U$ tal que: $|A| = \{k \wedge \forall u \in A, v \in (U - A) : d(q, u) \leq d(q, v)\}$

Las búsquedas por similitud pueden ser resueltas trivialmente por medio de una búsqueda exhaustiva, con una complejidad $O(n)$. Para evitar esta situación, se preprocesa la base de datos por medio de un algoritmo de indexación con el objetivo de construir una estructura de datos o índice, diseñada para ahorrar cálculos en el momento de resolver una búsqueda. El tiempo total de resolución de una búsqueda puede ser calculado de la siguiente manera: $T =$ evaluaciones de $d \times$ complejidad$(d) +$ tiempo extra de $CPU +$ tiempo de I/O.

En muchas aplicaciones la evaluación de la función d es tan costosa, que las demás componentes de la fórmula anterior pueden ser despreciadas. Este es el modelo que usaremos en este trabajo.

Básicamente existen dos enfoques para el diseño de algoritmos de indexación en espacios métricos: uno está basado en Diagramas de Voronoi [5, 2, 7] y el otro está basado en pivotes [5, 3, 4].

Uno de los principales obstáculos en el diseño de buenas técnicas de indexación es lo que se conoce con el nombre de *maldición de la dimensionalidad*. El concepto de dimensionalidad está relacionado con el nivel de dificultad al buscar en un determinado espacio métrico. La dimensión de un espacio métrico se define en [5] como $\rho = \frac{\mu^2}{2\sigma^2}$, siendo μ y σ^2 la media y la varianza respectivamente de su histograma de distancias. Es decir que, a medida que la dimensionalidad intrínseca crece, la media crece y su varianza se reduce. Esto significa que el histograma de distancia se concentra más alrededor de su media, lo que influye negativamente en los algoritmos de indexación.

3 SiBDaNN: Un Sistema de Bases de Datos para la Identificación de NN's

En este trabajo abordamos la aplicación de la teoría de Espacios Métricos para la identificación de cadáveres en el contexto de búsqueda de personas desaparecidas. El objetivo es desarrollar un sistema que permita mantener una base de datos, modelizada con un espacio métrico, con información sobre cadáveres no identificados para posteriormente realizar búsquedas de personas desaparecidas. Por cada cadáver se mantendrá un vector con los datos de las características físicas del mismo. Al momento de realizar una búsqueda, se deberá ingresar el vector con los datos de las características físicas de la persona buscada para que el sistema realice una búsqueda por similitud sobre la base de datos de cadáveres. El resultado será una lista de cadáveres con características similares a la persona buscada, rankeados según el grado de similitud.

Explicamos a continuación el trabajo desarrollado hasta el momento.

3.1 Generación de Vectores Característicos

Para la elaboración del sistema, como primer paso hubo que definir un core de datos que ses adecuado a la problemática de identificación. Usar pocos datos podría provocar que las búsquedas que se realicen sean de muy baja selectividad y usar demasiados puede provocar que se descarten elementos de la base de datos que sean de interés. En este sentido, ya hemos definido un primer core de datos sobre el cual trabajar: *color de ojos, color de pelo, color de piel, existencia de tatuajes y lugar de los mismos, cicatrices y/o marcas* (lunares por ejemplo), si existen *amputaciones y en qué lugar del cuerpo*, la *contextura física (atlético, atrófico, etc.)* y finalmente la existencia de *agenesias*. Para cada dato, se transforman los valores de su dominio en números que reflejen el grado de similitud entre los valores considerados y se genera el vector correspondiente.

Claramente no todos los datos tienen el mismo grado de importancia, por ejemplo el color de pelo es menos importante que el color de ojos porque una persona puede cambiarse el color de pelo pero no el de ojos. Esto hay que tenerlo en cuenta para establecer para cada dato un peso que corresponda con el grado de importancia del mismo.

3.2 Función de Distancia

Otro punto importante es la función de distancia a utilizar en las búsquedas. En una primera etapa usaremos la función coseno [1]. Si q es la query (vector de la persona buscada) y d_j es el *j-ésimo* vector de la base de datos, entonces el grado de similitud entre los vectores $\boldsymbol{d_j}$ y $\boldsymbol{q}$ se calcula como el coseno del ángulo formado entre ambos vectores:

$$sim(d_j, q) = \frac{\boldsymbol{d_j}.\boldsymbol{q}}{|\boldsymbol{d_j}| \times |\boldsymbol{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}$$

donde w_{iq} es el peso del *i-ésimo* dato en la consulta q y w_{ij} es el peso del *i-ésimo* dato en el vector d_j .

3.3 Indexación y Búsquedas

Con respecto al algoritmo de indexación comenzaremos usando algoritmos basados en pivotes. Cuando la base de datos se cargue con datos reales, se podrá analizar la dimensionalidad del espacio métrico sobre el cual se está trabajando y de ser necesario se cambiará el algoritmo de indexación.

Con respecto a las búsquedas, utilizaremos las búsquedas de los k vecinos más cercanos, porque es la que más se adecúa a este problema. Esto permitirá al usuario del sistema decidir cuánto elementos desea recuperar en una primera instancia y luego, de ser necesario, podrá ampliar la búsqueda; por ejemplo: puede pedir los 10 cadáveres mas parecidos a la persona buscada y posteriormente puede ampliar la búsqueda pidiendo los 10 siguientes.

4 Conclusiones y Trabajo Futuro

En este trabajo abordaremos la problemática de identificación de cadáveres en el contexto de búsqueda de personas con el fin de dar un primer paso que sirva de ayuda a una problemática tan delicada.

Para el desarrollo del sistema usamos el modelo de espacios métricos para la realización de búsquedas por similitud. Hasta el momento se ha programado el front-end del sitio web que permitirá la conexión con el sistema SiBDaNN. Como trabajo futuro nos proponemos programar el back-end e iniciar la prueba del sistema. En función de los resultados que se obtengan con la primer versión del sistema se realizarán, de ser necesario, cambios que pueden implicar: aumentar o disminuir los datos del core, cambiar la función de distancia y/o cambiar el algoritmo de indexación.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley, 1999.
2. S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.
3. W. Burkhard and R. Keller. Some approaches to best-match file searching. *Comm. of the ACM*, 16(4):230–236, 1973.
4. E. Chávez, J. Marroquín, and G. Navarro. Fixed queries array: A fast and economical data structure for proximity searching. *Multimedia Tools and Applications (MTAP)*, 14(2):113–135, 2001.
5. E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.
6. Procuraduría de Trata y Explotación de Personas (PROTEX) y la Colectiva de Intervención antes las Violencias (CIAV). Búsqueda de personas en democracia: Identificaciones de nn, trayectorias de vidas y cursos burocráticos. Technical report, Ministerio Público Fiscal, 2020.
7. G. Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.